1) A) (i) Label : 1

(ii)

		Predicted +	Predicted −
Correct +	+	TP	FN
	−	FP	TN

		Predicted +	Predicted −
Correct	+	209	64
	−	134	93

(iii) $TP\ Rate = \frac{\#TP}{\#TP + \#FN} = \frac{209}{209+64}$

$= \frac{209}{273} = 0.766$

$FP\ Rate = \frac{\#FP}{\#FP + \#TN} = \frac{134}{134+93}$

$= \frac{134}{227} = 0.590$

Accuracy = 100 × (Correct Predicted / Total Predicted) = (209 + 93 / 500) × 100

= (0.604) × 100 = 60.4 %.

(iv) Label : 1

(v)

		Predicted	
		+	−
Correct	+	212	61
	−	136	91

(vi) TP Rate = #TP / #TP + #FN = 212 / 212 + 61

= 212 / 273 = 0.778

FP Rate = #FP / #FP + #TN = 136 / 136 + 91 = 136 / 227

= 0.599

$$\text{Accuracy} = 100 \times \frac{(\text{Correct Predicted})}{(\text{Total Predicted})} = 100 \times \frac{(212+91)}{500}$$

$$= 100 \times \frac{303}{500} = 0.606 \times 100 = 60.6\%.$$

viii #)

		Predicted	
		+	−
Correct	+	273	0
	−	227	0

(b) Distance function that we are using would not be a good choice because in this we are just looking at the unique tokens between 2 comment as a measure for distance & then dividing it by 1.

But this would give rise to certain problems

1) Case 1:- When C_1 is small as compared to C_2. let C_1 & C_2 both belongs to +ve class.

Even if they both belong to same class they would be at a long distance from each other. because they would only have few tokens in common. As compared to other example in which C_1 & C_2 are of same length. So C_2 would not be predicted as +ve.

Case 2: When C_1 & C_2 are both of large length. Let C_1 belongs to -ve class & C_2 to +ve.

Then distance between them would be less as they would have a lot of common tokens (like punctuation, 'am', 'is', etc)

in Common. So C_2 ~~they~~ would be predicted as 'no' even when it is 'yes'.

2) We are dividing 1 by the unique tokens in common in C_1 & C_2. So when they both have large Number of tokens in common. So dividing that by 1 would make it pretty small. So a lot of examples would be ~~stated~~ ~~near~~ calculated to be near each other as the difference would be in some small decimal Number between them. If the difference is in only at certain decimal point like 0.00178493631 & 0.00178936 30 then it would be hard for us to predict accurately because we store number as an approximation in variables.

(i) Accuracy for $K=3$ is 65.9%.

Accuracy for $K=7$ is 68.8%.

Accuracy for $K=9$ is 61.2%.

(ii) Accuracy on test dat set with $K=3$ is 59%.

We have used $K=3$ because it has Maximum Accuracy on training.

		Predicted	
		+	−
Correct	+	212	61
	−	144	83

d (i) Distance function = $\frac{1}{\text{Uncommon unique word in } C_1 \& C_2}$

Uncommon unique words means only rare or less common words that is we remove the words like 'the', 'he', 'she', 'are', 'is', 'at' etc. as well as punctuation marks like '.', ':', ';' etc.

For eg.
C_1 = "Hello, Its me. You look nice."
C_2 = "nice of you., The weather."

Hence uncommon = {nice}
which is in C_1 and C_2

It would remove tokens like 'you', '.', ',', Because they are common words, and hence don't influence the positivity or negativity of a comment.

(ii) In first, the distance function, we calculate the number of unique words that are common in both C_1 and C_2. But, when we do that we count tokens like '!', '.' 'and' etc. even when they have no influence on the positivity or negativity of a comment.
So when we remove these common tokens we are left with exactly the number of token which are influential. Thus, the token will have a greater influence on the comment being positive or negative.

(iii) Confusion Matrix

		Predicted	
		+	−
Correct	+	TP 230	FN 43
	−	FP 88	TN 139

(IV) True Positive Rate = $\frac{\#TP}{\#TP + \#FN}$

$= \frac{230}{230+43} = \frac{230}{273}$

$= 0.84$

False positive Rate = $\frac{\#FP}{\#FP + \#TN}$

$= \frac{88}{88+139}$

$= 0.388$

Accuracy = $\frac{TP + TN}{Total}$

$= \frac{230 + 139}{500}$

$= 0.738$ or 73.8%

(V) Confusion Matrix for $K=5$

		Predicted +	Predicted −
Correct	+	235	38
	−	94	133

vi) True Positive Rate = $\frac{\#TP}{\#TP + \#FN} = \frac{238}{235 + 38}$

$= \frac{238}{273} = 0.872$

False Positive Rate = $\frac{\#FP}{\#FP + \#TN} = \frac{94}{94 + 133}$

$= \frac{94}{227} = 0.414$

Accuracy = $\left(\frac{TP + TN}{Total}\right) \times 100 = \left(\frac{235 + 133}{500}\right) \times 100 = 0.736 \times 100$

$= 73.6\%$

(vii) Yes, it did achieves higher accuracy both for $K = 1$ and $K = 5$

For $K = 1$

Old function accuracy = 60.4 %

New distance function accuracy = 73.8 %

For $k = 5$

Old distance function accuracy = 60.6%
New distance function accuracy = 73.6%

Done